基于用户画像分析的媒体知识库的设计与实现

马 鸣 陈辛夷 陈 珺
（新华通讯社通信技术局技术研发中心，北京 100083）

摘 要：互联网、大数据和新媒体技术的发展带来媒体传播渠道和内容形态革新性变化，分析用户在不同渠道新闻采用和传播情况是挖掘用户行为、构建用户画像、增强稿件传播力和采用率的重要组成部分。然而，由于新媒体技术的发展，媒体融合成为趋势，用户和媒体的关系变得更加复杂多变，传统的媒体统计服务方式难以满足使用者对用户采用情况的直观化、可视化、多维化的需求。本文从多源头收集用户、机构、媒体等单位信息，利用大数据、规则匹配、标签提取等技术，构建了基于用户画像分析的媒体知识库，充分发掘了数据资源的要素潜力，打破了系统项目间的信息壁垒，推动了数据协同管理和服务，让高质量的数据资源为更多的研发产品提供有力的数据支持和保障。文中介绍了基于用户画像分析的媒体知识库的建设意义和设计原理，对关键技术和实现进行了深入研究，在此基础上提出了基于媒体知识库的智能化应用展望。

关键词：媒体知识库；媒体融合；用户画像；统计采用 **中图分类号：**G206 **文献标识码：**A
文章编号：1671-0134（2022）03-145-04 DOI：10.19483/j.cnki.11-4653/n.2022.03.046
本文著录格式：马鸣，陈辛夷，陈珺. 基于用户画像分析的媒体知识库的设计与实现 [J]. 中国传媒科技，2022（03）：145-148.

导语

根据中国互联网络信息中心 2021 年发布的中国互联网发展统计报告 [1]，截至 2020 年 12 月，我国网民规模达 9.89 亿，普及率达到 70.4%。随着互联网市场规模的不断扩大，越来越多的新型媒体开始出现。面对信息化时代向大数据时代的转型期，传统的媒体统计服务方式难以满足使用者对媒体采用的直观化、可视化、多维化的需求。同时，由于新媒体技术的发展，媒体融合成为趋势，用户和媒体的关系变得更加复杂多变。微博、微信、抖音、快手等多种社交媒体开始成为新闻信息传播的途径。为了更快捷地传递新闻信息，扩大自身影响力，一大批传统媒体建起了自己的媒体矩阵，形成了“一点多翼”的传播格局。

为了充分发掘数据资源要素潜力，打破传统系统间的信息壁垒，更好发挥数据的基础资源作用和创新引擎作用，本文构建了一种以媒体矩阵为核心的知识库。同时，为了有效利用大数据技术和核心数据资源，分析用户在不同渠道的新闻采用和传播情况，发掘传播媒体的深度及广度，维护品牌形象、提升影响力，本文介绍了基于用户画像分析的媒体知识库的建设意义和设计原理，对关键技术和实现进行了深入研究，在此基础上提出了基于媒体知识库的智能化应用展望。

1. 构建基于用户画像分析的媒体知识库的必要性和可行性

1.1 构建基于用户画像分析的媒体知识库的必要性

建立用户画像模型，能够为稿件精准化推送、提升传播影响力提供决策参考。而随着互联网以及移动互联技术的高速发展，媒体在发展过程中进入到快速融合的时代。信息的签发渠道在一定程度上发生改变，使得受众得以分流，同时受众对于单一渠道的关注不断的下降，用户与媒体的关系变得更加复杂。[2] 因此，为了建立用户画像模型、整合数据资源，积极响应媒体深度融合，本文提出了一种构建基于用户画像分析的媒体知识库模型。将用户纳入融合体系，才能真正构建全程、全息、全员、全效的融媒体平台，在媒体竞争中赢得自己的格局空间。

1.2 构建基于用户画像分析的媒体知识库的可行性

首先，经过近几年的建设，统计监测系统已积累起数亿条稿件落地传播数据，采集的媒体家次已达到 33 万 + 家次，同时随着供稿系统的建设，积累了大量的用户订阅数据。尽管不同系统项目间存在信息壁垒，但大量结构化数据为构建媒体知识库提供了数据支持。其次，媒体融合的大趋势，逐渐形成行业经验。王一佼 [3] 以“华理日报”为例，搭建了新媒体矩阵，介绍了多平台运营，详述了新媒体矩阵的用户反馈。陈杏兰 [4] 在介绍媒体融合实例的同时，提出了“媒体矩阵”建设中的三个思维误区，强调用户在媒体融合中的大作用。刘静等人 [5] 从政务新媒体矩阵的建设出发，介绍了在政务系统中，新媒体矩阵的构成及发展困境。这些实践为构建基于用户画像的媒体知识库提供了丰富的建模方法和应用案例。

2. 构建基于用户画像分析的媒体知识库的技术原理

2.1 媒体知识库数据的获取

媒体知识库意图构建基于用户 - 机构 - 媒体三种属性的媒体知识矩阵。其中，用户数据来源于新华社通信技术局供稿系统的注册用户。机构数据来源于第三方资源引入、大数据技术爬取和人工整理。媒体数据来源于第三方资源引入、互联网采集资源、统计监测系统的媒

体库。下面将介绍机构和媒体的数据获取和数据属性。

概念 1：机构，机关、团体或其他工作单位。

概念 2：媒体，传播信息的媒介，包括传统媒体和新媒体，属于机构，根据渠道具有不同的标识属性。

本文前期对 110 万机构名称和 480 万公司名称的数据集 [6] 进行了数据清洗、整理和关键字聚类等工作，经数据清洗后，共得到与新闻相关的机构名称 19701 家和与新闻相关的公司名称 68153 家，总计 87854 家，为后续完善媒体知识库的机构表提供数据支持。其中与新闻相关的机构关键字聚类结果分布如图 1。

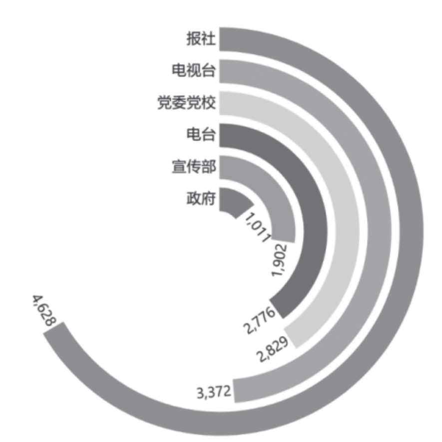


图 1　关键字聚类结果分布

媒体数据在类型上分为传统媒体和新型媒体。传统媒体如报纸、电视、广播、网站等，在新闻信息的传递和采用上依旧占据较大比例。同时，随着用户接收信息方式的改变，“两微一端”，抖音，快手等短视频平台、B 站、腾讯视频等长视频平台，逐渐成为舆论的主战场、主阵地。

本文整理了国内和海外社交媒体中占据主流、影响力较大的媒体渠道，为后续根据渠道采集互联网媒体信息、引入拓尔思第三方资源提供样例支持，主要媒体渠道如图 2 所示。



图 2　主要媒体渠道

2.2　媒体知识库的维度

构建基于用户画像的媒体知识库的核心是要对用户、机构、媒体三方面的数据标签化，从不同的维度构建标签体系，使媒体知识库更加具体、更加可靠、更加全面。

媒体知识库的主要指标根据用户、机构、媒体三种不同的概念而设计了三种不同的指标体系。在机构方面，主要考察机构的基础信息、地域属性和影响力属性。其中基础信息主要记录了机构的权威信息，地域属性记录了机构的地区分布，影响力属性以成立年限、中央 / 地方、商业 / 非商业等标签构成，旨在后续对机构的分析更加全面。在媒体方面，主要考察媒体的基础信息、地域属性和渠道属性。媒体的渠道属性根据网站、“两微一端”、海外社交平台、短视频平台、长视频平台、公众号等社交平台的不同，建立了统一的媒体唯一标识并提取了各渠道的渠道信息。在用户方面，主要考察用户的基础信息和订阅信息。媒体知识库的主要指标信息如表 1 所示。

构建媒体知识库，就是要构建媒体矩阵，建立用户—机构—媒体三者的关联关系，发现机构和媒体间的关系，完善机构的下属媒体及该媒体渠道的信息，为后续建立用户画像、进行个性化推荐提供数据支持。

构建基于用户画像的媒体知识库的本质是将用户—机构—媒体数据充分利用，根据媒体矩阵，将用户的需求用可视化的方式展现出来，应用于统计采用的服务中，辅助采编决策，实现精准化服务。基于用户画像的媒体库模型构建可分为三层——数据层、分析层、应用层，如图 3 所示。

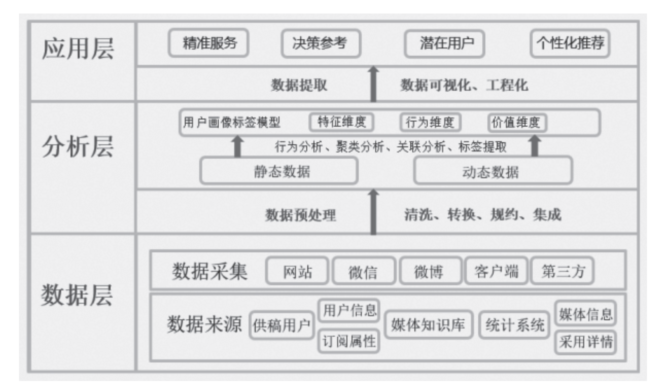


图 3　基于用户画像的媒体知识库框架

数据层是构建媒体知识库的基础层，分为数据来源和数据采集。数据来源主要从供稿系统、统计监测系统、第三方资源引入等获取；数据采集主要从互联网资源引入项目、县级融媒体采集小组处获得网站、微信、微博、客户和其他第三方媒体的采用数据，并将采集到的媒体数据序列化后存储到原始数据库中。并根据媒体入库情况，及时地对原始数据库中的数据进行更新与完善，建立用户—机构—媒体的关联关系。最后，通过对采集和引入的数据进行数据清洗、数据转换、数据规约、数据集成，为分析层做进一步的分析做准备。

分析层是根据媒体知识库用户—机构—媒体的关联

表 1　主要信息示意表

名称	一级信息	二级信息
机构	基础信息	机构名称
		机构别名
		外文名称
	区域属性	所属国家
		所属省份
		所属城市
		所属县级
		所属境域
		所属大洲
	影响力属性	所属性质
		所属级别
		所属资质
媒体	基础信息	媒体名称
		外文名称
	区域属性	所属国家
		所属省份
		所属城市
		…
	渠道及信息	媒体渠道
		媒体账号
		媒体语种
		媒体类型
		媒体域名 / 网站
		所属机构
		…
用户	名称	用户名称
	订阅信息	订阅状态
		订阅详情

关系，将机构—媒体的地域、资质等静态数据和发放频率、评论阅读、采用等动态数据相结合，利用行为分析、聚类分析、关联分析、标签提取等技术建立媒体知识库的标签体系，实现用户特征标签化。通过标签建模分析，可以进一步挖掘出用户个体特征和群体特征，为应用层的个性化智慧服务提供支持。

应用层是在数据层和分析层工作的基础上提出基于媒体知识库的个性化服务。旨为利用大数据分析技术为采编人员提供决策参考，优化线路栏目；为用户提供精准化服务，个性化推荐；为营销人员提供版权维护、挖掘潜在用户。

2.3　媒体知识库的实现

2.3.1　建立数据同步机制，存储结构化数据

除了引入第三方资源，媒体知识库建立了与供稿系统、统计监测系统的数据同步机制。每日更新用户的订阅状态和入库媒体数据，及时完善用户—机构—媒体的关联关系，保证媒体知识库的时效性。

2.3.2　大数据赋能匹配，减少人工处理

前期通过对 590 万的实体数据进行数据清洗、聚类分析，得到共 8 万多家的机构名称数据。但 8 万 + 的机构名称数据不能直接使用。由于 590 万的数据是通过分词算法基于大数据得到的实体词，因此实体词的判断取决于分词算法的准确性。尽管如此，8 万 + 机构名称数据仍具有重要的参考价值。一方面，数据结果涵盖了媒体行业的方方面面，提供了较全的境内机构和公司名称清单。另一方面，机构和公司的命名相对较规范，同一机构或公司的名称叫法相对较全。因此，通过匹配规则程序上实现部分用户和机构关联关系的建立。实现程序自动匹配为主，人工调整为辅的用户 – 机构关系的建立模式。

2.3.3　提供关联关系查询，满足用户画像需求

构建基于用户画像分析的媒体知识库，就是充分利用大数据技术、规则匹配技术，建立全面、完善的用户—机构—媒体的标签体系。因此，在吸收了互联网资源引入、统计监测系统、供稿系统、第三方资源引入等多种数据后，媒体知识库作为独立系统对外提供查询服务，以实现媒体信息查询、机构信息查询、媒体矩阵查询、用户—机构—媒体的用户采用查询等多种查询方式，为后续用户画像、版权检测、稿件推荐等智能化应用提供数据支持。

3. 媒体知识库应用——以县级融媒体中心用户为例

为了加快县级融媒体中心融合，积极响应中央提出的加强县级融媒体的建设要求，本文收集并建立了全国 328 家县级融媒体中心的媒体矩阵，共生成 1241 家媒体。平均每家县级融媒体中心包含 4 种传播渠道。结合县级融媒体中心用户、媒体知识库和统计系统的 2020 年稿件采用情况，初步形成对县级融媒体中心用户的用户分析。

3.1　用户画像

构建用户画像模型，就是对用户的偏好及习惯利用媒体的订阅信息、采用记录形成标签，大量用户媒体标签的集合形成一个具有相同特征的用户群，为综合媒体知识库开展有针对性的媒体偏好活动提供了有效保障。下面从 4 点分析县级融媒体中心用户。

一是县级融媒体中心用户的传播渠道以“两微一端”为主，抖音、快手等短视频平台较少，B 站、腾讯视频等长视频平台的账号运营仍有不足。

二是结合稿件的采用情况，在采用稿件类型中，仍以图文类稿件为主，图文类稿件具有信息量大、存储量级小等优势，对视频类稿件的采用较少，这表明一方面采编人员要加强媒体融合能力，将图文类、音频类稿件与视频制作相结合，另一方面融媒体中心要加强平台运营能力，增加新闻传播形式，吸引受众。

三是在栏目排行详情中，“城乡发展”栏目的采用率最高，达到了 80% 以上，其次是“总书记报道”“时政新闻”栏目，采用率皆在 70% 以上。排名较靠后的栏目为“体坛快报”“健康百科”“财经聚焦”栏目。这表明相比于体育、财经等主题，县级融媒体中心用户更

偏爱与当地相关的时政新闻。

四是县级融媒体中心的区域属性。在用户采用排行TOP50中，属于内蒙古、甘肃、北京地区的县级融媒体中心用户排名较高。这表明在上述三个地区中，一方面县级融媒体中心发展较早，建立了相对较为完善的媒体融合播发机制，另一方面表明相比于其他区域，新华社在上述三个地区的推广较大、影响力较高。

3.2 个性化推荐

通过构建用户画像，发掘用户行为习惯，掌握用户用稿偏好，针对不同用户团体做个性化推荐。根据上述分析的栏目排行和区域属性，为用户做推荐时优先“城乡发展”栏目、县级融媒体专线线路和当地稿件，为增强用户黏性提供有效的技术支持。

3.3 版权监测

为版权监测系统提供用户采用情况、机构采用情况、媒体采用情况，通过大数据提取、云识别、信息比对等技术手段，对稿件作品或稿件数字化信息进行版权监测，有助于及时发现和查处侵权盗版行为，有助于查找侵权盗版线索、获取有力的电子证据，有效保护版权。同时，还可通过上述的用户画像和个性化推荐，为用户提供创作资料、偏好稿件，对提高新闻传播、提升社内影响力具有一定的推动作用。

4. 总结

媒体知识库建立了用户与媒体的关联关系，从用户数据入手，收集并汇总全球范围内的机构及其各渠道的子属媒体，构建了有层次、多渠道的媒体矩阵知识库。系统亮点在于系统完善扩充了媒体渠道，提供了机构的媒体矩阵，建立了自动更新机制，具有稳定的用户、机构、媒体三方面的数据来源。同时，建立了数据思维，打破了系统间的壁垒，实现了高效迅捷的信息查询方式，为后续基于媒体知识库的智能化应用提供数据支持。

参考文献

[1] CNNIC. 中国互联网络发展状况统计报告 [R]. 北京：中国互联网络信息中心，2018.

[2] 戴健允 . 融媒体时代打造媒体矩阵传播力 [J]. 中国宽带，2021（7）：191.

[3] 王一佼 . 出版业私域流量运营初探 ——以“华理日语”新媒体矩阵为例 [J]. 现代出版，2021（2）：85-88.

[4] 陈杏兰 . “媒体矩阵”建设中的三个思维误区 [J]. 传媒，2020（11）：65-67.

[5] 刘静，凌以民 . 我国政务新媒体矩阵的建设分析 [J]. 出版广角，2020（19）：23-25.

[6]https：//github.com/nirenxiaoxiao/Company-Names-Corpus

作者简介：马鸣（1993-），女，河北邯郸，硕士，工程师，研究方向：算法应用；陈辛夷（1986-），女，籍贯福建，硕士研究生，工程师，研究方向：新闻行业大数据；陈珺（1977-），女，四川，硕士研究生，高级工程师，研究方向：数据分析。

（责任编辑：张晓婧）

（上接第58页）

短视频。

4.4 新闻事件的时效性

媒体融合大环境之下，新闻传播方式逐步趋向于高效性。通过利用多种渠道报道新闻事件，就能够引起社会大众的关注，因此，新闻信息具备高效性的特征，新闻信息传输速度也会提升。在新闻发布之后，公众们就可以利用媒体平台发表自己的观点，在网络上，各网友评论也会进一步扩大该事件的影响力，使得新闻不仅仅局限于屏幕上。

4.5 传播平台多样化

在进入5G信息时代，新旧媒体融合，传统媒体与互联网诞生的新媒体都可以传递信息。使新闻传播具有多样化的特征。

结语

在当前5G技术发展的新环境，要进一步推动媒体融合，科学利用5G技术，迎接未来的技术发展变革。

参考文献

[1] 徐蕾 .5G 背景下媒体融合发展浅析 [J]. 出版广角，2020（1）：38-40.

[2] 李海军 .5G 时代媒体融合发展对策研究 [J]. 中国广播电视学刊，2019（5）：49-52.

[3] 张明新，常明芝 .5G 应用背景下媒体融合发展的前景 [J]. 新闻爱好者，2019（8）：12-14.

[4] 刘悦 .5G 时代媒体融合发展与创新策略研究 [J]. 中国有线电视，2020（5）：46-47.

作者简介：高平（1998-），女，重庆，硕士，研究方向：中国文化境外传播。

（责任编辑：张晓婧）